

# Modelação Ecológica

## AULA 13

29 October 2019 – 14:00-16:30 – room 2.3.37


Tiago A. Marques





RE: Estágios 2019-2020...


File Message Help PDFsam Enhanced 5 Creator Tell me

Delete Respond Quick Steps Move Tags Editing Speech Zoom View Headers

## RE: Estágios 2019-2020

 Miguel Mascarenhas <miguel.m@bioinsigt>  
To: Tiago Marques; Helena Coelho Wed 1:32 PM

 You replied to this message on 10/23/2019 5:06 PM.

Oi Tiago,

Se houver alunos interessados em fazer mini-estágios, tipo 1 semana ou 4 semanas ou 1 trabalho a efetuar durante 1 mês mas que podem ir fazendo a partir de casa (tipo alguma modelação) é uma questão de virem falar connosco e preparamos um estágio personalizado 😊

Abraço, Miguel.

---

**De:** Tiago Marques <[tiago.marques@st-andrews.ac.uk](mailto:tiago.marques@st-andrews.ac.uk)>  
**Enviado:** 23 de outubro de 2019 10:42  
**Para:** Helena Coelho <[helena.c@bioinsight.pt](mailto:helena.c@bioinsight.pt)>  
**Cc:** Miguel Mascarenhas <[miguel.m@bioinsight.pt](mailto:miguel.m@bioinsight.pt)>



← Tweet

Home

Explore

Notifications

Messages

Bookmarks

Lists

Profile

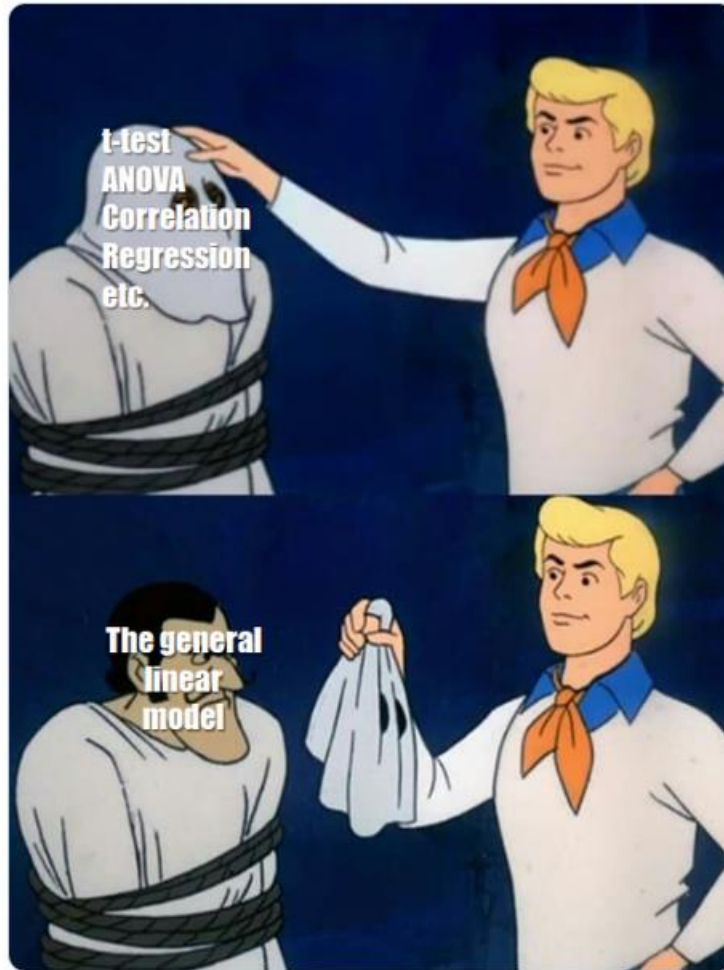
More

Tweet



Stuart Ritchie  
@StuartJRitchie

I made this meme for our stats class last week and I thought you might like to see it.



11:54 AM · Oct 27, 2019 · Twitter Web App

370 Retweets 1.9K Likes



Search Twitter

### Relevant people



Stuart Ritchie  
@StuartJRitchie

Follow

Lecturer at @SGDPCCentreKCL, King's College London. Looks like a 'cartoonish' 'startled hedgehog'.

### Trends for you



Trending in Portugal

**Halloween**

1.39M Tweets

Trending in Portugal

**Queen**

267K Tweets

Trending in Portugal

**Star Wars**

65K Tweets

Trending in Portugal

**Malta**

11.9K Tweets

Trending in Portugal

**Argentina**

739K Tweets

Show more

Terms Privacy policy Cookies Ads info

More © 2019 Twitter, Inc.

# PROGRAMME

09:00-09:10 *Welcome words* (L. Carriço, Dean)

09:10-09:25 *Facts and figures about research @ CIÊNCIAS* (M Santos-Reis, Vice Dean for Research)

## SESSION I - Top Notch Science

09:25-09:40 *Out of this world atmospheres* (Pedro Machado)

09:40-09:55 *Active matter* (Nuno Araújo)

09:55-10:10 *The 1755 earthquake and the closing of the Atlantic Ocean* (João C. Duarte)

10:10-10:25 *Transcutaneous electric stimulation of the spinal cord: a modelling study* (Pedro C. Miranda)

## 10:25-11:00 Coffee-break

11:00-11:15 *Glycofighting bacteria: a new mode of action* (Rodrigo Almeida)

11:15-11:30 *A new mechanism to inhibit amyloid aggregation in Alzheimer's Disease* (Cláudio M. Gomes)

11:30-11:45 *How Mediterranean and Tropical forests react to groundwater change?* (Cristina Antunes)

11:45-12:00 *Vulnerability & Blame: making sense of unauthorized access to smartphones* (Diogo Marques)

## 12:00-14:30 Bring a sandwich, look at the posters and have a speed date

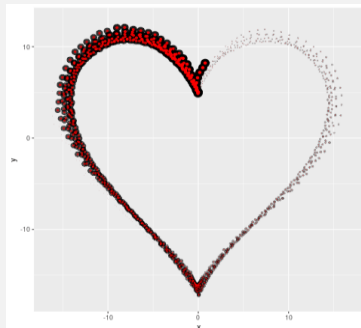
*Speed dating the (great) experts behind great scientists!* (12:30-13:30 - C3 Building, Atrium)

*Speed dating a statistician* (13:00-14:00 - C3 Building, Atrium)

*Speed dating mathematicians* (13:00-14:30 - C6 Building, Room 6.1.8)

*THE LAST M\_{(e)}LE\_{(l)}* (13:00-14:30 - C6 Building, Room 6.1.8)

(More info about the speed dating sessions [available here](#)).



@accidental\_aRt by @csmarcum



Sociedade / Eventos /

## Ciências Research Day - programme now available

Share (0) Tweet (0) LinkedIn (0)

Out	Localização
30	Edifício C3, FCUL, Lisboa
2019	

A melhor Ciência faz-se em CIÊNCIAS!

## SESSION II - Recognising Excellence (ERC grantees)

14:30-14:40 *Why this, why now, why me?* (Joaquim Gaspar)

14:40-14:50 *Competition under (niche) construction: an ERC project (not so) easy to construct* (Sara Magalhães)

14:50-15:00 *Where's Wally?: Spotting the next ERC grantees at CIÊNCIAS* (Henrique Leitão)

## SESSION III - Networking and Science for Society

15:00-15:15 *Intelligent infection management and precision antibiotherapy* (Ricardo Dias)

15:15-15:30 *Estimating the efficacy of mass rescue operations in ocean areas with vehicle routing models and heuristics* (Rui de Deus)

15:30-15:45 *CoastNet - Portuguese Coastal Monitoring Network* (José L. Costa)

15:45-16:00 *SmartHub Energy* (Miguel Brito)

## 16:00-16:30 Coffee Break

16:30-16:45 *Ciências at the core of European efforts to push the boundary of physics* (António Amorim)

16:45-17:00 *Making the added value of networking tangible* (Raquel Conceição)

17:00-17:15 *The Art of spinning-off* (Fadhil Musa)

## SESSION IV - Challenging Ideas for Ciências: Creative Minds Contest

17:15-17:45 *Pitch talks*

17:45-18:00 *Closing remarks and Awards* (Pedro Almeida, Vice Dean for Communication and Image)



Ciências  
ULisboa

Tomorrow... we will be celebrating  
*Ciências Research Day* – so... no class!



É já no dia 30 de outubro que se vai realizar o *Ciências Research Day*, a 1.<sup>a</sup> edição daquele que é o maior evento sobre a investigação científica que se faz em Ciências ULisboa.

Participe em todas as dimensões do programa e conheça o trabalho desenvolvido por docentes e investigadores da sua Faculdade - esta é por isso uma oportunidade de criar novas colaborações e projetos.

**The reason there's no class is so that you can come and participate in the day... if not (in fact, even if you come)... you are expected to use the time to work in your ME assignments: MECOCO, the theoretical R package work or the final practical work ...**

**...so many options, so little time!**

# A COOL PACKAGE FOR AUTOMATED EXPLORATORY DATA ANALYSIS

```
library(dlookr)
texugo <- read.csv2("Texugo.csv")
eda_report(texugo, target = Densidade, output_format = "html", output_file = "EDA.html",
output_dir=getwd())
```

response variable

Puts doc in the working directory

data.frame with the data

## Exploratory Data Analysis Report

Report by dlookr package

2019-10-27

- 1 Introduction
  - 1.1 Information of Dataset
  - 1.2 Information of Variables
  - 1.3 About EDA Report
- 2 Univariate Analysis
  - 2.1 Descriptive Statistics
  - 2.2 Normality Test of Numerical Variables
    - 2.2.1 Statistics and Visualization of (Sample) Data
- 3 Relationship Between Variables
  - 3.1 Correlation Coefficient
    - 3.1.1 Correlation Coefficient by Variable Combination

# Exploratory Data Analysis Report

Report by dlookr package

2019-10-27

- 1 Introduction
  - 1.1 Information of Dataset
  - 1.2 Information of Variables
  - 1.3 About EDA Report
- 2 Univariate Analysis
  - 2.1 Descriptive Statistics
  - 2.2 Normality Test of Numerical Variables
    - 2.2.1 Statistics and Visualization of (Sample) Data
- 3 Relationship Between Variables
  - 3.1 Correlation Coefficient
    - 3.1.1 Correlation Coefficient by Variable Combination

The dataset that generated the EDA Report is an **'data.frame'** object. It consists of **14 observations** and **15 variables**.

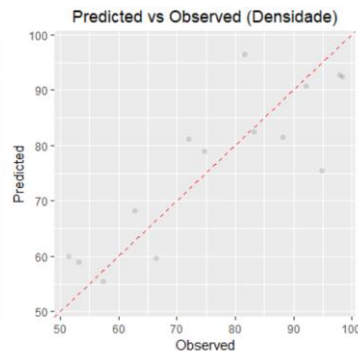
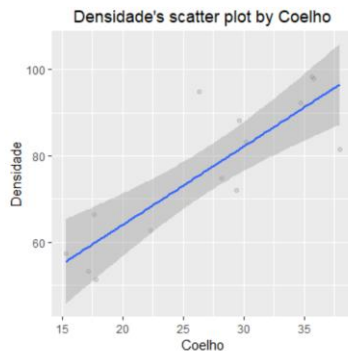
## 1.2 Information of Variables

The variable information of the data set that generated the EDA Report is shown in the following table:

Information of Variables

variables	types	missing_count	missing_percent	unique_count	unique_rate
Densidade	numeric	0	0	14	1.0000000
Percentagem.floresta	numeric	0	0	12	0.8571429
Coelho	numeric	0	0	14	1.0000000
Raposa	numeric	0	0	13	0.9285714

Visualização:



## 2.2 Normality Test of Numerical Variables

### 2.2.1 Statistics and Visualization of (Sample) Data

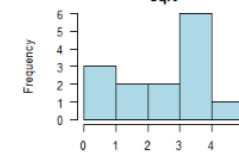
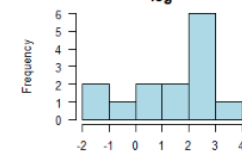
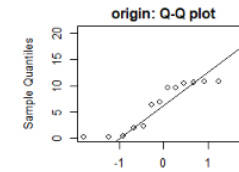
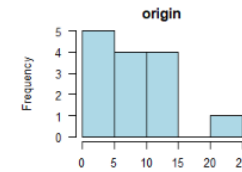
[ **Percentagem.floresta** ]

normality test : Shapiro-Wilk normality test

statistic : 0.88657, p-value : 0.0721477

skewness and kurtosis

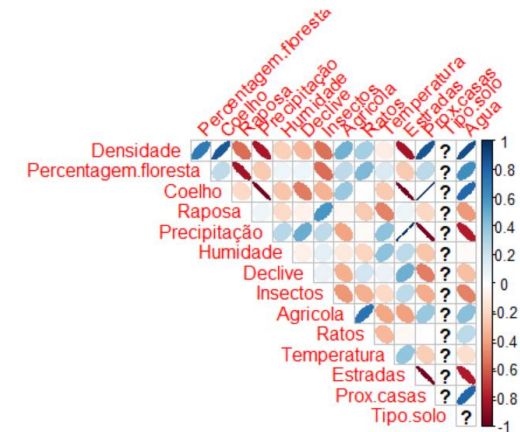
type	skewness	kurtosis
original	0.5451584	3.010174
log transformation	-0.9635486	2.412474
sqrt transformation	-0.3083150	1.980650



[ **Coelho** ]

normality test : Shapiro-Wilk norr

### 3.1.2 Correlation Plot of Numerical Variables



# AND... THE FOLLOWING WEEK

Tuesday 5<sup>th</sup> November

Théo Michelot: an introduction to HMMs in Ecology, with an emphasis on animal movement (a really nice opportunity to interact with a leader in his field - a course inside our course – bonus: a participation certificate will be provided, so your CV will look nicer!)

Wednesday 6<sup>th</sup> November

Students will do their own research – I am available for questions as usual.

Your mission will be to use GLMs to model some datasets, including a gamma regression and a beta regression – data sets will be on FENIX by the 1<sup>st</sup> November 2019.

You can also work in your MECOCO assignment, or the final work, or the theoretical work... so many options, so little time!



# GUIÃO PARA O TRABALHO PRÁTICO

## Gestão de Páginas

- ▼ Moderação Ecológica
  - Modelação Ecológica(Ecologia Marinha)
  - Modelação Ecológica(Ecologia e Gestão Ambiental)
- ▶ Aulas
- ▼ Outros Recursos
  - ▶ PDFs
    - R Cheat Sheets
    - Propostas de resolução de fichas de trabalho
    - Bioinsight
  - ▼ Avaliação
    - MECOCO 20 % individual
    - Practical 45 %**

+ Criar

## Practical 45 %

Página **Ficheiros 1** Permissões Link

Adicionar Ficheiro

#	Nome
1	GuiãoTrabalhoPratico.pdf GuiãoTrabalhoPratico.pdf

# Generalized Linear Models (continued!)



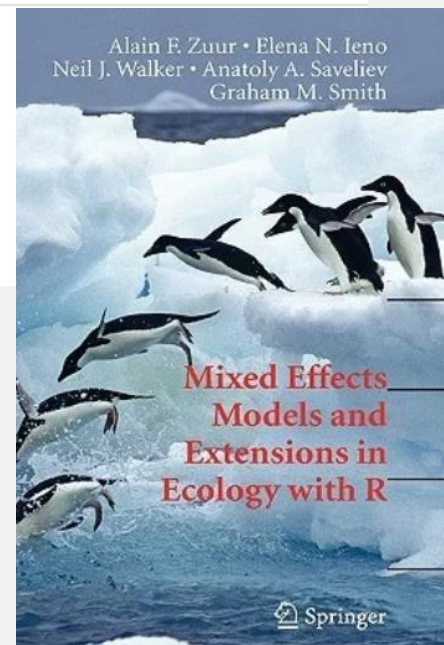
Sunday, May 14, 2017

A gentle introduction to Generalized Linear Models in R

What are generalized linear models?

<http://r-eco-evo.blogspot.com/2017/05/generalized-linear-models.html>

<http://spatialecology.weebly.com/r-code--data/category/glm>



# Residuals in a GLM context

Standard residuals:  $y_i - \hat{y}_i$  a.k.a.  $y_i - \hat{\mu}_i$

## Pearson residuals

$$\hat{\varepsilon}_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(Y_i)}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

A standardized Pearson residual is obtained by dividing the Pearson residual by the  $\sqrt{1 - h_i}$  where  $h_i$  is the leverage of observation  $i$

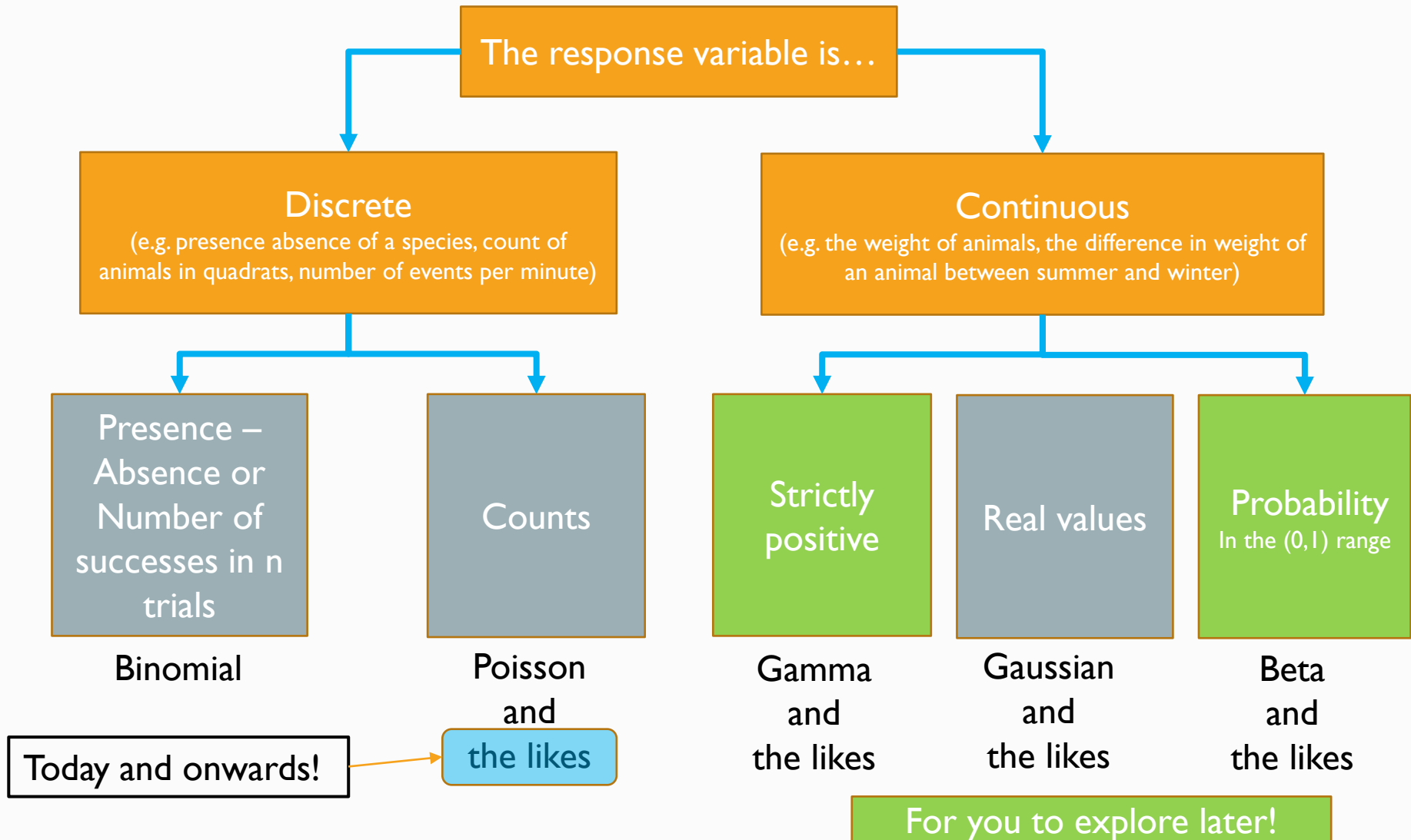
## Deviance residuals

$$\hat{\varepsilon}_i^D = \text{sign}(y_i - \mu_i) \sqrt{d_i}$$

The  $d_i$  represents the contribution of the  $i^{\text{th}}$  observation to the deviance

Reminder: The notion of Gaussian residuals does not apply to GLMs

When modelling a response variable, what we want is a model that describes the data, eventually as a function of covariates. Therefore, the first step is to decide what will be the distribution of the response variable. In other words, what to use in the family argument of most modelling functions in R (like `glm`, `gam`, etc.). There are a couple of big questions that need answering:



# Modeling counts:

when the Poisson is not enough!

*Quasi-stuff*, Negative Binomial, Zero  
Inflated models & Mixture models,  
Truncated Models

Quasi-*stuff*

An alternative to deal with **overdispersion** is to give up on the Poisson, and move towards an option where there is not a specific distribution for the response variable, but



an assumed relationship between the mean and the variance

### *9.7.3 Quick Fix: Dealing with Overdispersion in a Poisson GLM*

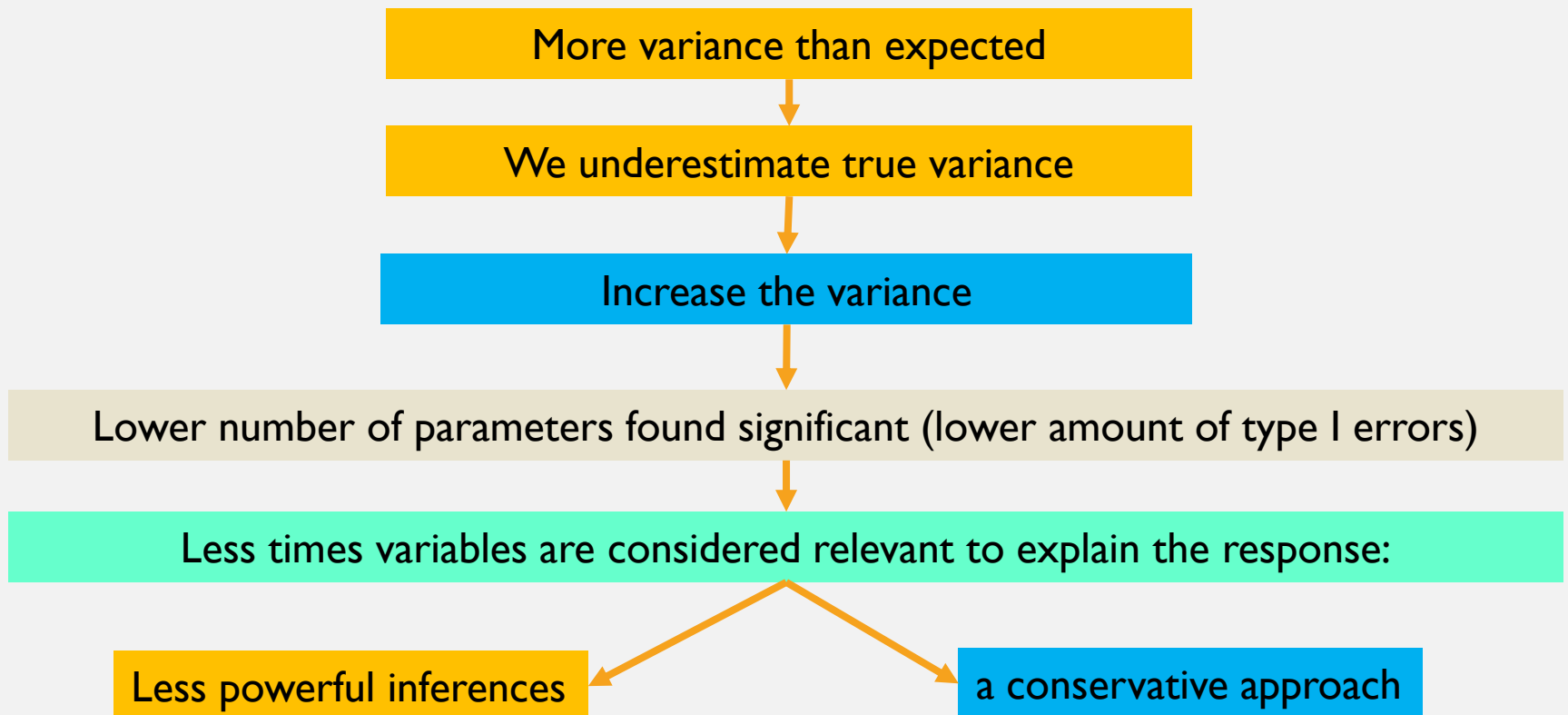
We can deal with overdispersion in the GLM by using a quasi-Poisson GLM, which consists of the following steps:

1. The mean and variance of  $Y_i$  are given by  $E(Y_i) = \mu_i$  and  $\text{var}(Y_i) = \phi \times \mu_i$ .
2. The systematic part is given by  $\eta(X_{i1}, \dots, X_{iq}) = \alpha + \beta_1 \times X_{i1} + \dots + \beta_q \times X_{iq}$ .
3. There is a logarithmic link between the mean of  $Y_i$  and the predictor function  $\eta(X_{i1}, \dots, X_{iq})$ .

The direct consequence of doing this is that the standard errors (se) of parameters will be increased proportionally to the dispersion parameter.

$$se(\text{parameter}_{\text{quasi}}) \approx \sqrt{\phi} se(\text{parameter})$$

Larger standard errors is the correct thing to do





```

##
## Call:
## glm(formula = species.richness ~ dist.coast, family = quasipoisson,
##      data = data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.4753  -3.0008  -1.7069   0.9003   6.4290
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.894e+00  2.152e-01   8.800 3.57e-13 ***
## dist.coast  4.135e-06  1.510e-06   2.738 0.00772 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.31074)
##
##      Null deviance: 768.25  on 76  degrees of freedom
## Residual deviance: 694.62  on 75  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5

```

```
##
## Call:
## glm(formula = species.richness ~ dist.coast, family = poisson(link = "log"),
##      data = data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.4753  -3.0008  -1.7069   0.9003   6.4290
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.894e+00  6.702e-02  28.258  <2e-16 ***
## dist.coast  4.135e-06  4.703e-07   8.792  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 768.25  on 76  degrees of freedom
## Residual deviance: 694.62  on 75  degrees of freedom
## AIC: 975.72
##
## Number of Fisher Scoring iterations: 5
```

and explanatory variables. It is a software issue to call this 'quasipoisson'. Do not write in your report or paper that you used a quasi-Poisson distribution. Just say that you did a Poisson GLM, detected overdispersion, and corrected the standard errors using a quasi-GLM model where the variance is given by  $\phi \times \mu$ , where  $\mu$  is the mean and  $\phi$  the dispersion parameter. To get the numerical output for this model,

Quasi

No likelihood means no derived measures

~~AIC~~

How to make model selection under quasi-likelihoods?



# Model selection options

```
drop1(glmQP1, test="F")
```

→ This is not a “FALSE”, but the “F-test”!

```
## Single term deletions
##
## Model:
## species.richness ~ dist.coast
##           Df Deviance F value  Pr(>F)
## <none>          694.62
## dist.coast  1    768.25  7.9498 0.006148 **
```

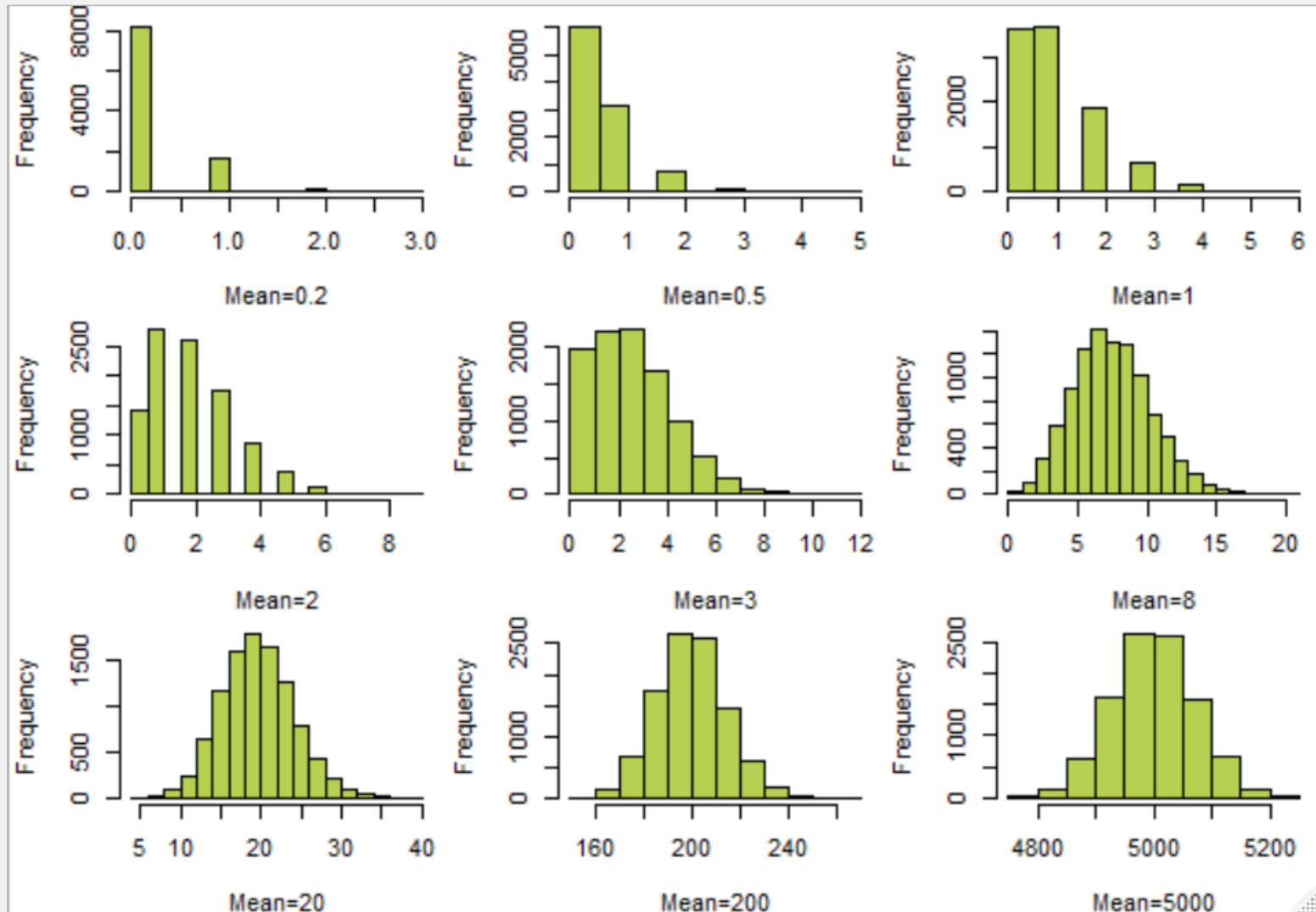
```
glmQP2=glm(species.richness~1, family=quasipoisson, data=data)
anova(glmQP1, glmQP2, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: species.richness ~ dist.coast
## Model 2: species.richness ~ 1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         75     694.62
## 2         76     768.25 -1  -73.628 0.007534 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Negative Binomial

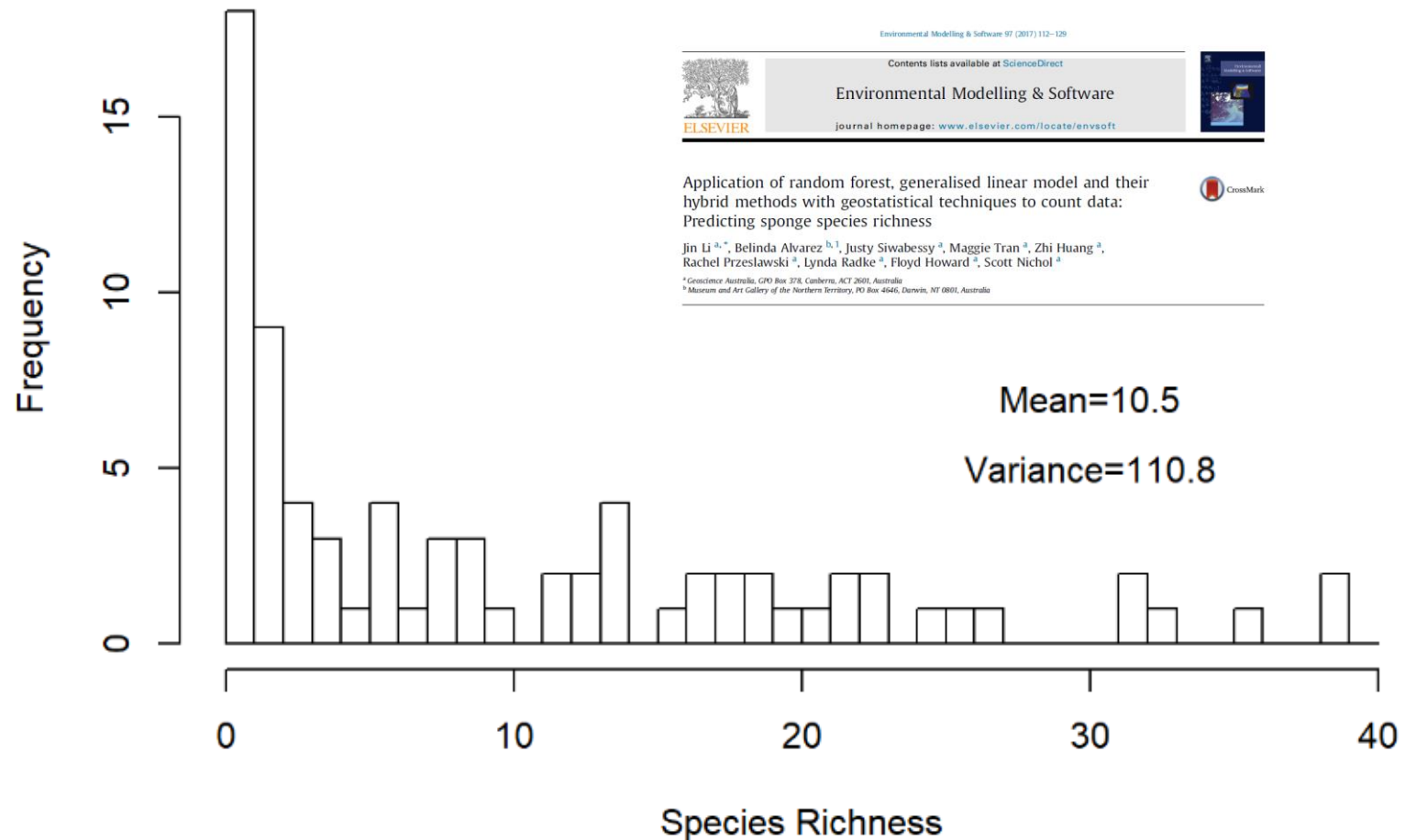
Not just adjusting standard errors  
but  
truly accounting for the overdispersion

The Poisson, the default distribution for counts, looks like this



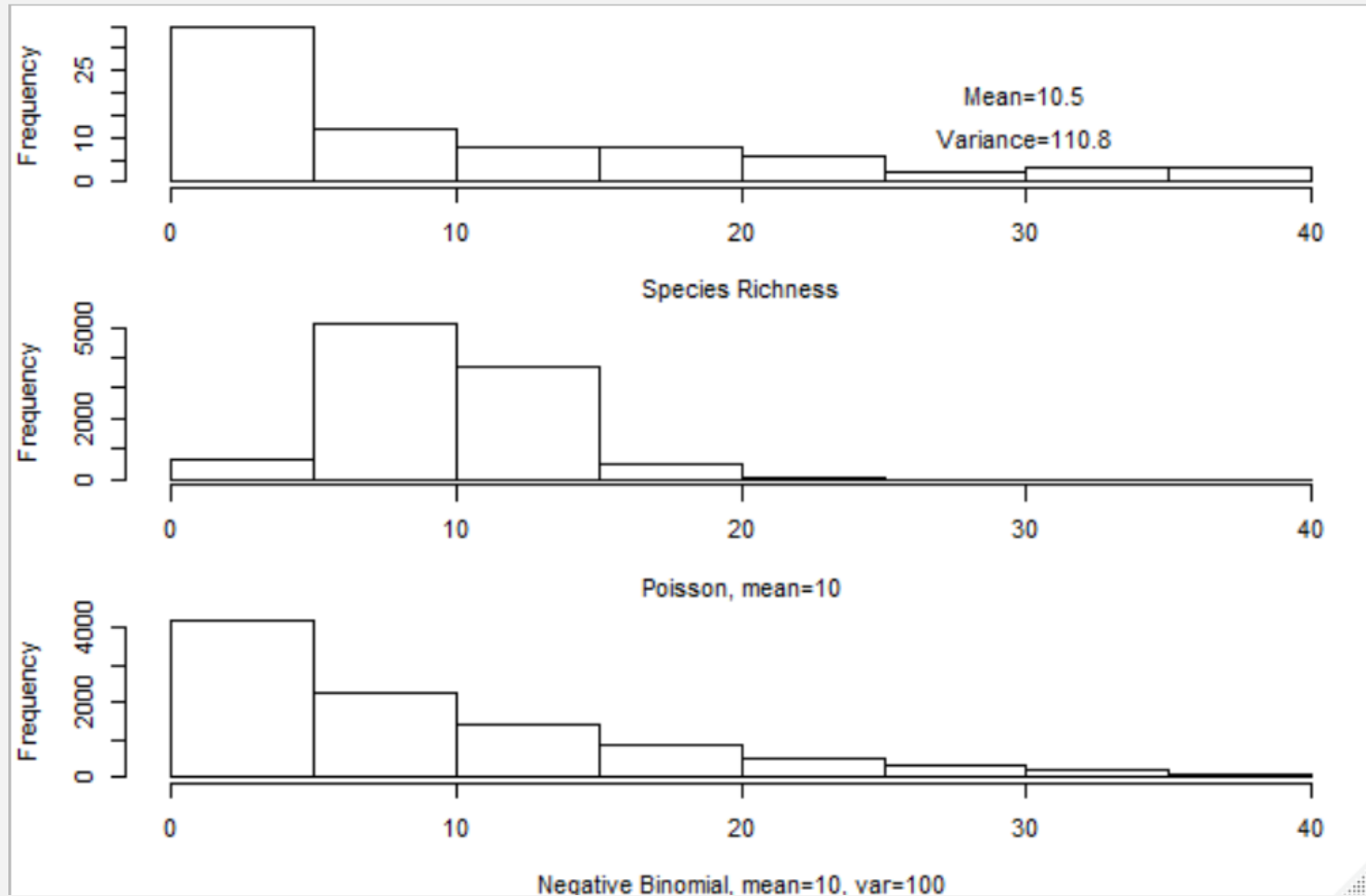
The **mean = variance** is a strong, **often unreasonable**, assumption

Most real ecological count datasets present **overdispersion** (i.e. mean  $>$  variance) compared to the Poisson



Clearly, the **variance** is much larger than the mean!

Most real ecological count datasets present **overdispersion** (i.e. mean  $>$  variance) compared to the Poisson



Here, but also often for real data, the **negative binomial** seems a priori a much better choice!



$$Y_i \sim NB(\mu_i, k)$$

$$E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \mu_i + \frac{\mu_i^2}{k}$$

Overdispersion component (what is “extra” compared to the mean)

```
> #testing a random variable!!  
> glmP1=glm(species.richness~dist.coast,family=poisson(link="log"),data=data)  
> library(MASS)  
> glmNB1=glm.nb(species.richness~dist.coast,link=log,data=data)  
> AIC(glmP1,glmNB1)
```

Function to fit  
Negative Binomial GLM

	df	AIC
glmP1	2	975.7237
glmNB1	3	522.3367

```

## Call:
## glm.nb(formula = species.richness ~ dist.coast, data = data,
##       link = log, init.theta = 1.084393054)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5856  -1.2639  -0.5590   0.2566   1.7763
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.871e+00  2.052e-01   9.117 < 2e-16 ***
## dist.coast   4.345e-06  1.669e-06   2.604  0.00923 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0844) family taken to be 1)
##
##      Null deviance: 89.908  on 76  degrees of freedom
## Residual deviance: 83.032  on 75  degrees of freedom
## AIC: 522.34
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  1.084
##              Std. Err.:  0.184

```

There is still a slight  
overdispersion component

# Zero Inflated models & Mixture models

Ecological data, quite often, presents a much larger amount of 0's than what usual standard statistical distributions (e.g. Poisson, Negative Binomial, etc), can cope with

Zero inflated data – a real issue to model adequately

*Ecology Letters*, (2005) **8**: 1235–1246

doi: 10.1111/j.1461-0248.2005.00

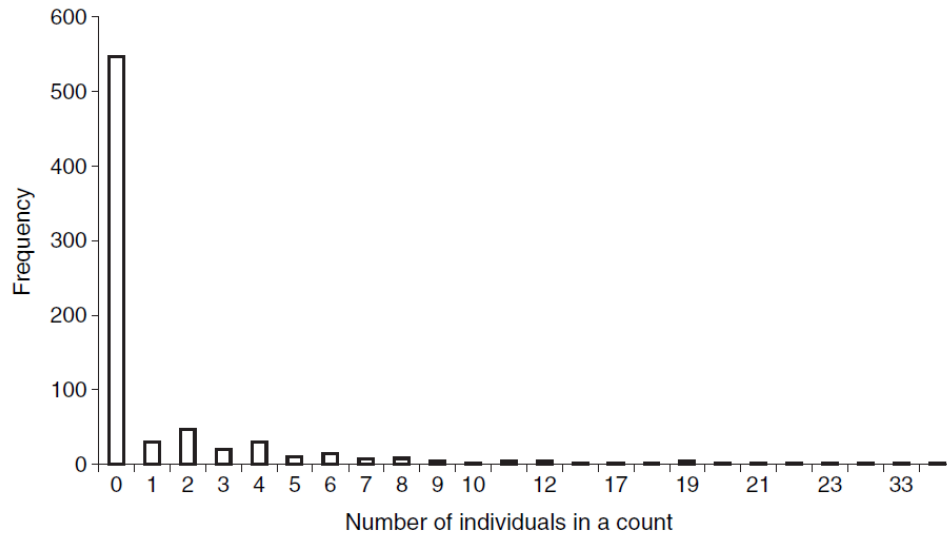
**REVIEWS AND  
SYNTHESES**

**Zero tolerance ecology: improving ecological inference by modelling the source of zero observations**

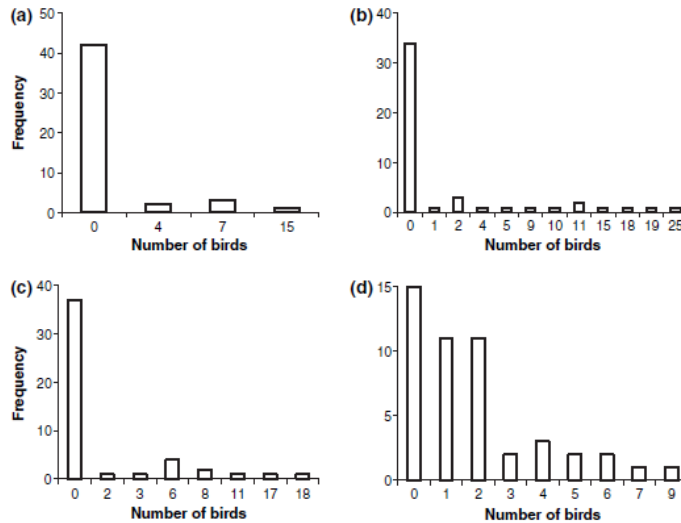
Tara G. Martin,<sup>1\*</sup> Brendan A. Wintle,<sup>2</sup> Jonathan R. Rhodes,<sup>3</sup> Petra M. Kuhnert,<sup>4</sup> Scott A. Field,<sup>5</sup> Samantha J. Low-Choy,<sup>6</sup> Andrew J. Tyre<sup>7†</sup> and Hugh P. Possingham<sup>1</sup>

**Abstract**

A common feature of ecological data sets is their tendency to contain many zero values. Statistical inference based on such data are likely to be inefficient or wrong unless careful thought is given to how these zeros arose and how best to model them. In this paper we propose a framework for understanding how zero-inflated data sets originate and how best to model them. We define and classify the different kinds of zero-inflation that occur in ecological data and describe how they arise: either from 'true zero' or




**Figure 1** Example of a typical zero-inflated data set. Frequency of counts for 31 bird species across eight sites and three grazing treatments ( $n = 744$ ) from Martin *et al.* (2005). Over 70% of the data set is represented by zero counts, which is more than expected if a Poisson distribution is assumed for the species' abundances.




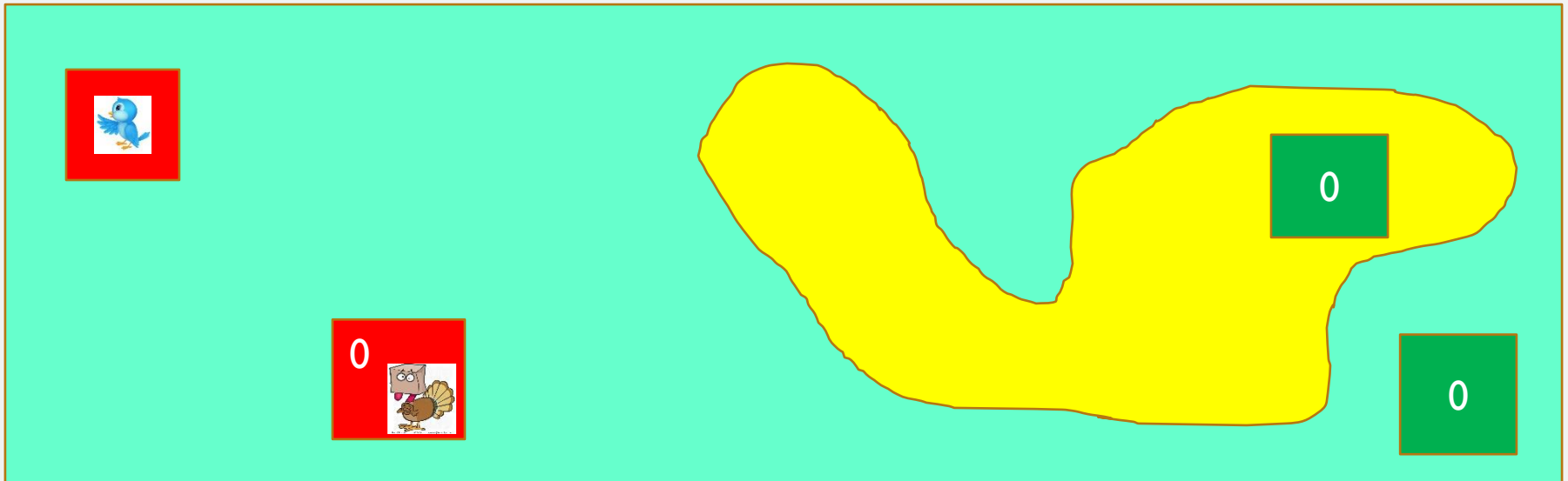
**Figure 2** Frequency of counts of four woodland bird species: (a) brown thornbill, (b) noisy miner, (c) superb fairy-wren and (d) rufous whistler across 24 sites visited twice in summer and twice winter.

**True zero** - Species does not occur at a site because of the ecological process, or effect under study (e.g. **habitat unsuitable**)

**True zero** - Species does not saturate its **entire suitable habitat** by chance

**False zero** - Species occurs at a site, but is not present during the survey period 

**False zero** - Species occurs at a site and is present during the survey period, but the observer fails to detect it (particularly common for rare or cryptic species )

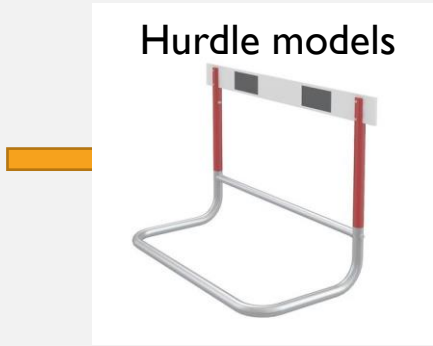


**Table 2** Four scenarios of zero occurrences in ecological data and the modelling approach recommended for presence/absence and count data, where zero inflation can be caused by false zeros, true zeros or a combination of both

Zero inflation	Modelling approach	Key references
None	Single distribution models (e.g. binomial)	McCullagh & Nelder (1989)
True zeros	Zero-inflated mixture models, ZIB or ZIP with point mass at zero, or hurdle models	Lambert (1992), Welsh <i>et al.</i> (1996) and Hall (2000)
False zeros	Zero-inflated mixture models (e.g. ZIB or ZIP)	MacKenzie <i>et al.</i> (2002, 2003) and Tyre <i>et al.</i> (2003)
Both	Mixture of two or more distributions	None found

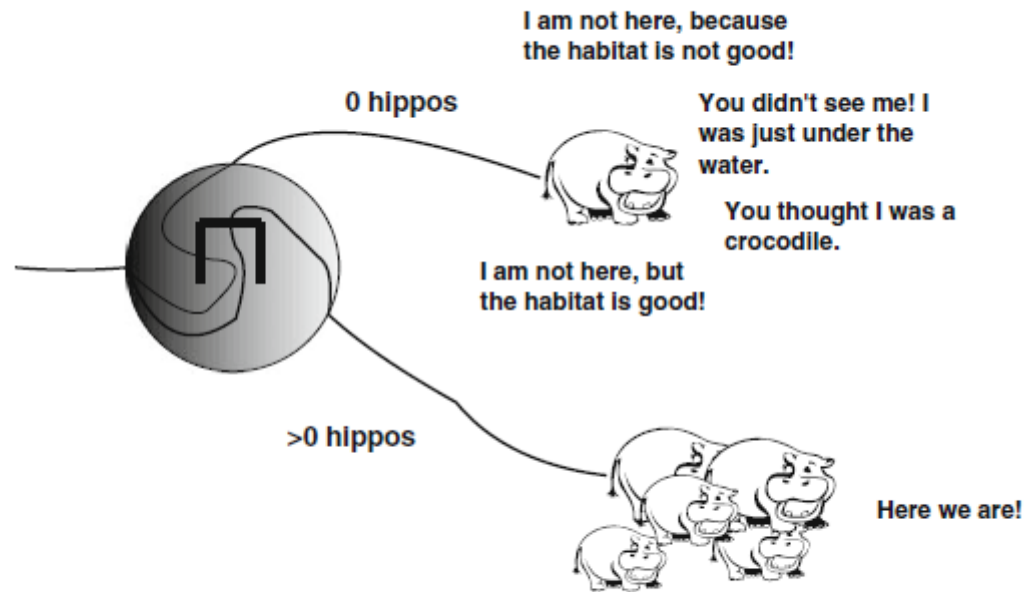
The zero-inflated models are based on the binomial distribution for presence/absence data, and on the Poisson or negative-binomial model for count data. ZIP, zero-inflated Poisson; ZIB, zero-inflated binomial.

Model the 0's and 1's

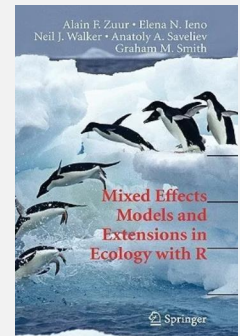


Conditional on not being a 0, model the count with a truncated distribution

# Hurdle models



**Fig. 11.4** Sketch of a two-part, or hurdle model. There are two processes; one is causing zeros versus non-zeros, the other process is explaining the non-zero counts. This is expressed with the hurdle in the *circle*; you have to cross it to get non-zero counts. The model does not make a distinction between the different types of zeros





```
#an example
```

```
f1=y~X1+X2+X3|X3+X4
```

```
H1A <- hurdle(f1, dist = "poisson", link = "logit", data = dados)
```

```
H1B <- hurdle(f1, dist = "negbin", link = "logit", data = dados)
```

It seems like you can only specify the link for the binomial component!

## hurdle and

The function `zeroinfl` allows the following formulae specifications.

1.  $Y \sim X_1 + X_2$ . This is equivalent to:  $Y \sim X_1 + X_2 | 1$ .
2.  $Y \sim X_1 + X_2 | X_1 + X_2$
3.  $Y \sim X_1 + X_2 | Z_1 + Z_2$

Only the logistic model is a function of covariates

Each component is a function of different covariates

Both components a function of the same covariates

## Check out the example code in ?hurdle

```
> ## logit-poisson
> ## "art ~ ." is the same as "art ~ . | .", i.e.
> ## "art ~ fem + mar + kid5 + phd + ment | fem + mar + kid5 + phd + ment"
> fm_hp1 <- hurdle(art ~ ., data = bioChemists)
> summary(fm_hp1)
```

Call:  
hurdle(formula = art ~ ., data = bioChemists)

Pearson residuals:

	Min	1Q	Median	3Q	Max
	-2.4105	-0.8913	-0.2817	0.5530	7.0324

Count model coefficients (truncated poisson with log link):

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	0.67114	0.12246	5.481	4.24e-08	***
femWomen	-0.22858	0.06522	-3.505	0.000457	***
marMarried	0.09649	0.07283	1.325	0.185209	
kid5	-0.14219	0.04845	-2.934	0.003341	**
phd	-0.01273	0.03130	-0.407	0.684343	
ment	0.01875	0.00228	8.222	< 2e-16	***

Zero hurdle model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	0.23680	0.29552	0.801	0.4230	
femWomen	-0.25115	0.15911	-1.579	0.1144	
marMarried	0.32623	0.18082	1.804	0.0712	.
kid5	-0.28525	0.11113	-2.567	0.0103	*
phd	0.02222	0.07956	0.279	0.7800	
ment	0.08012	0.01302	6.155	7.52e-10	***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 12  
Log-likelihood: -1605 on 12 Df

This slide was NOT shown in class but is very relevant here:

```
> fm_hp1 <- hurdle(art~fem+mar | fem, data = bioChemists)
> summary(fm_hp1)
```

Call:

```
hurdle(formula = art ~ fem + mar | fem, data = bioChemists)
```

Pearson residuals:

Min	1Q	Median	3Q	Max
-1.1328	-1.0324	-0.3321	0.3764	10.2825

Count model coefficients (truncated poisson with log link):

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	0.847598	0.064084	13.226	< 2e-16	***
femWomen	-0.237351	0.064199	-3.697	0.000218	***
marMarried	0.008846	0.066944	0.132	0.894867	

Zero hurdle model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	0.9679	0.1007	9.609	<2e-16	***
femWomen	-0.2604	0.1445	-1.802	0.0715	.

---

signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

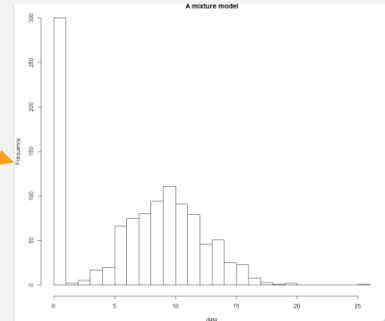
Number of iterations in BFGS optimization: 9

Log-likelihood: -1670 on 5 Df

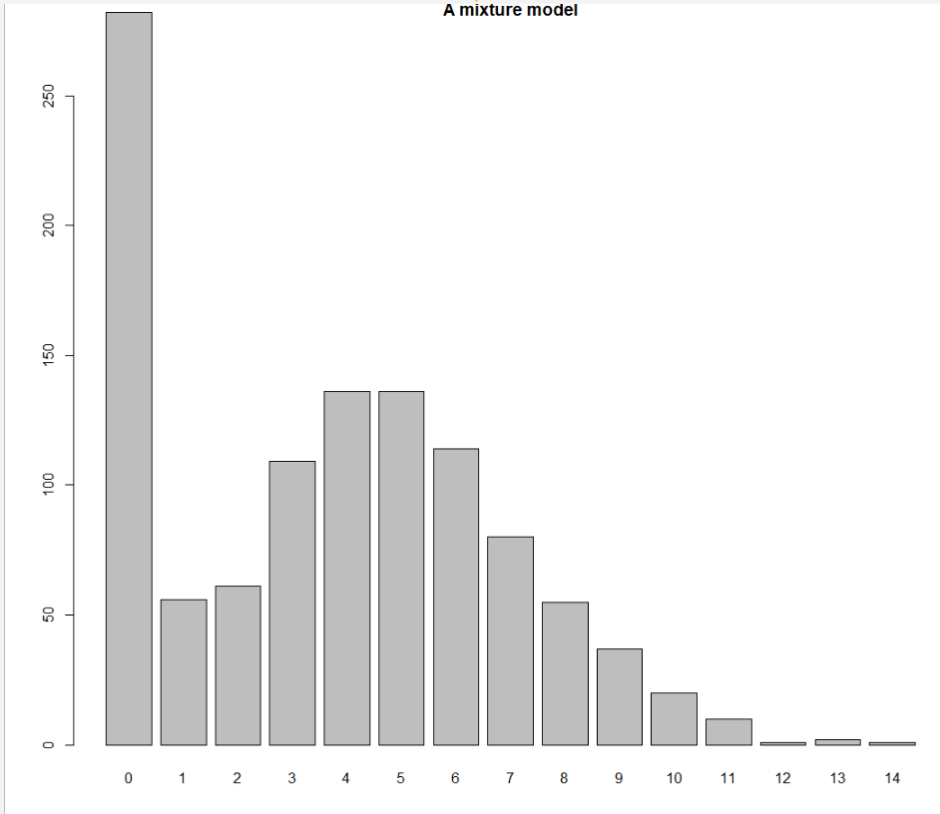
**Unlike what I said in class**, the **first component** in the formula is for the **count model**, while the **second component** is for the **binomial model**

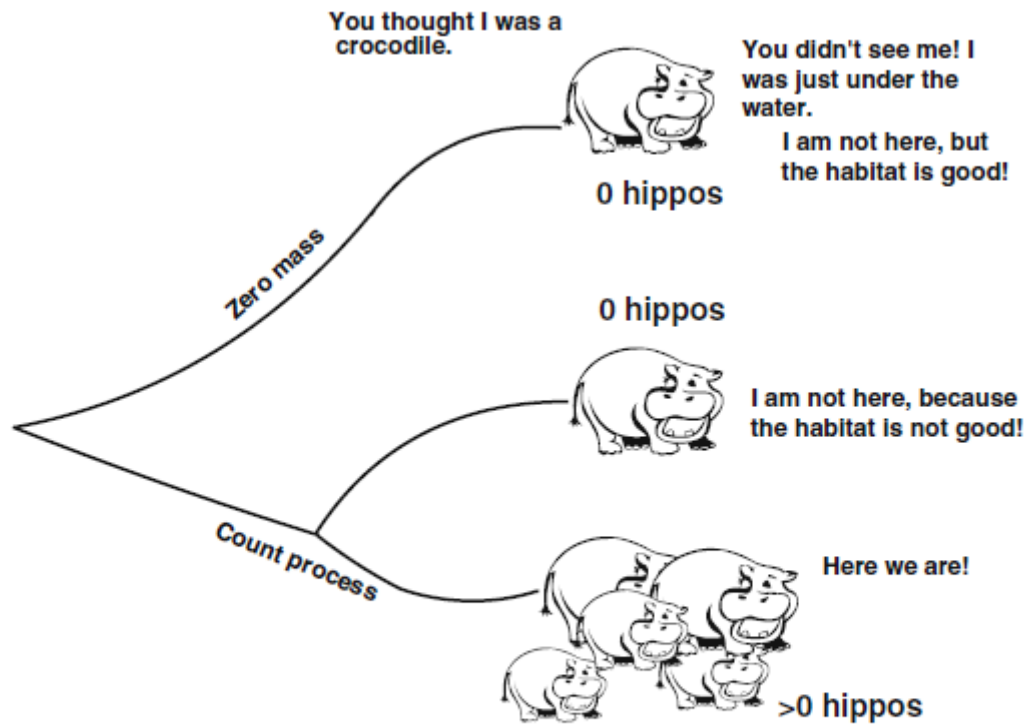
```
#a mixture model
set.seed(1234)
n=1000
ms=c(0.3,0.8)
p=0.1
meanP=5
zerosand1s=rbinom(n*ms[1],1,p)
counts=rpois(n*ms[2],meanP)
data=c(zerosand1s,counts)
par(mfrow=c(1,1),mar=c(4,4,0.5,0.5))
barplot(table(data),main="A mixture model")
```

meanP=10

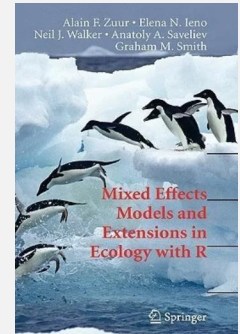


A mixture model





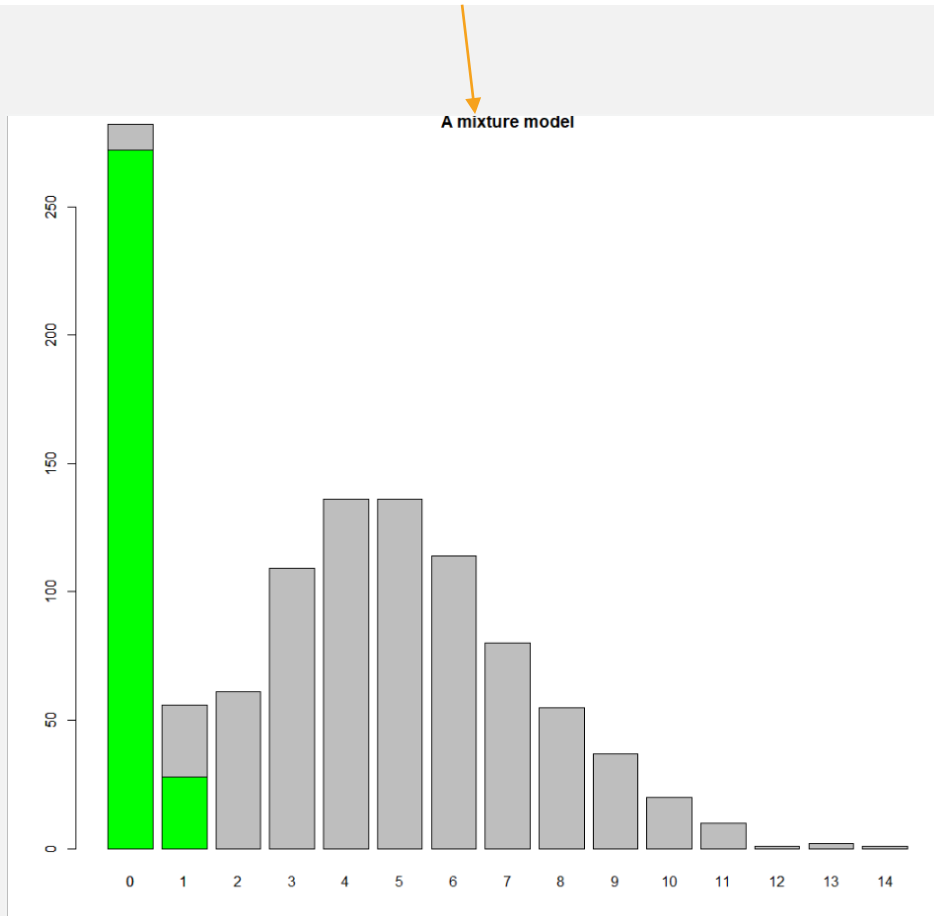
**Fig. 11.5** Sketch of the underlying principle of mixture models (ZIP and ZINB). In counting hippos at sites, one can measure a zero because the habitat is not good (the hippos don't like the covariates), or due to poor experimental design and inexperienced observers (or experienced observers but difficult to observe species)



```

#a mixture model
set.seed(1234)
n=1000
ms=c(0.3,0.8)
p=0.1
meanP=5
zerosandls=rbinom(n*ms[1],1,p)
counts=rpois(n*ms[2],meanP)
data=c(zerosandls,counts)
par(mfrow=c(1,1),mar=c(4,4,0.5,0.5))
barplot(table(data),main="A mixture model")
barplot(table(data[1:length(zerosandls)]),add=TRUE,col="green")

```



We could now try to explain in a regression model, using covariates, the two processes: one driving the 0's and the 1's, (the "green data") and one driving counts (including some 0's, the "grey data")

To implement these zero inflated (mixture) models we need additional packages. There are a few around, and this is just an example using library psc1

```
library(psc1)
#The zero inflated Poisson - mixture model
zip <- zeroinfl(y~X1+X2| X1+X2, dist = "poisson", link
= "logit", data = ParasiteCod2)
#The zero inflated Negative binomial - mixture model
zinb <- zeroinfl(y~X1+X2|X1+x3, dist = "negbin", link
= "logit", data = ParasiteCod2)
```

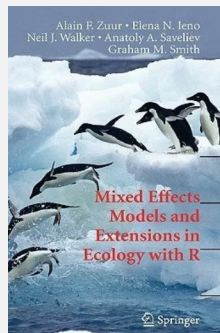
The function `zeroinfl` allows the following formulae specifications.

1.  $Y \sim X_1 + X_2$ . This is equivalent to:  $Y \sim X_1 + X_2 | 1$ .
2.  $Y \sim X_1 + X_2 | X_1 + X_2$
3.  $Y \sim X_1 + X_2 | Z_1 + Z_2$

Each component is a function of different covariates

Both components a function of the same covariates

Only the logistic model is a function of covariates



This slide was NOT shown in class but is very relevant here:

```
> fm_zinb2 <- zeroinfl(art ~ . |ment, data = biochemists, dist = "negbin")  
>  
> summary(fm_zinb2)
```

call:

```
zeroinfl(formula = art ~ . |ment, data = biochemists, dist = "negbin")
```

Pearson residuals:

Min	1Q	Median	3Q	Max
-1.3041	-0.7685	-0.2632	0.4671	6.3765

Count model coefficients (negbin with log link):					
	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	0.404004	0.141716	2.851	0.00436	**
femwomen	-0.211906	0.071922	-2.946	0.00322	**
marMarried	0.139462	0.081193	1.718	0.08586	.
kid5	-0.167624	0.052457	-3.195	0.00140	**
phd	0.001963	0.035586	0.055	0.95601	
ment	0.024393	0.003518	6.934	4.1e-12	***
Log(theta)	1.002890	0.142832	7.021	2.2e-12	***

Zero-inflation model coefficients (binomial with logit link):					
	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-0.8067	0.3532	-2.284	0.0224	*
ment	-0.6095	0.2458	-2.480	0.0131	*

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Theta = 2.7261

Number of iterations in BFGS optimization: 32

Log-likelihood: -1553 on 9 Df

**Unlike what I said in class**, the **first component** in the formula is for the **count model**, while the **second component** is for the **binomial model**



# Truncated Models

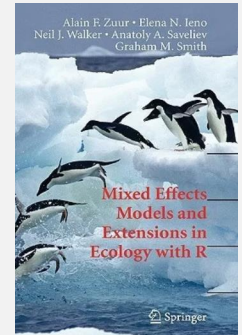
Not that common, but still!

Sometimes we might have situations in which some values are not possible for the response variable

As an example, say if numbers below or above a threshold are not possible

Under such scenario, one might consider **truncated** models

The most common example is when 0 is not a possibility (but in theory you could truncate any set of values from a distribution)



Pag. 268

Syntax for a zero truncated example

```
M3A <- vglm(N_days ~ PDayRain + Tot_Rain + Road_Loc +  
  PDayRain:Tot_Rain, family = posnegbinomial,  
  control = vglm.control(maxit = 100),  
  data = Snakes)
```

pospoisson is the Poisson counterpart!



ration. The group size in this modelling data ranged between 1 and 6 whales. We considered the following as potential explanatory variables: (1)  $K$ , the total number of the hydrophones over which the dive echolocation clicks were detected, (2)  $n$ , the total number of clicks detected across all hydrophones, (3)  $d$ , the duration of the echolocation period (time difference between the first click and the last click associated to the dive), and (4) the detected click rate ( $\frac{n}{d}$ ). Additionally, we considered variables that, while not being related to group size per se, could affect the detected acoustic footprint and hence obscure the relationship between the acoustic footprint and the group size if ignored. These were binary variables indicat-

## Estimating group size from acoustic footprint to improve Blainville's beaked whale abundance estimation

Tiago A. Marques<sup>b,\*</sup>, Patrícia A. Jorge<sup>a</sup>, Helena Mouriño<sup>a</sup>, Len Thomas<sup>c</sup>, David J. Moretti<sup>d</sup>, Karin Dolan<sup>d</sup>, Diane Claridge<sup>e</sup>, Charlotte Dunn<sup>e</sup>

**Table 1**

Models, variables considered (*cdur* = click duration, *nhyd* = number of hydrophones and *crate* = detected click rate), the corresponding coefficient value and respective P-value, as well as the Akaike's Information Criteria.

Model	Variable	Coefficient	P-value	AIC
M1	<i>crate</i>	0.002	0.030	150.55
M2	<i>nhyd</i>	-0.056	0.021	150.71
	<i>crate</i>	0.003	0.193	
M3	<i>cdur</i>	0.010	0.011	150.49
	<i>nhyd</i>	-0.093	0.071	
	<i>crate</i>	0.004	0.137	

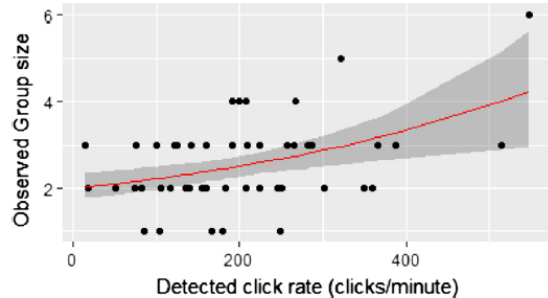


Fig. 1. Observed group sizes and corresponding detected click rate (black dots), along with the modelled relationship (red line), and the model's bootstrap 95% percentile interval for the mean group size (grey area). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

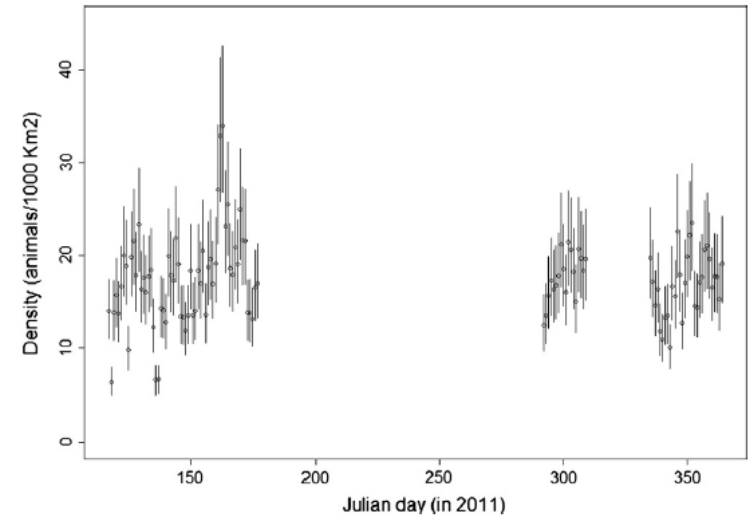


Fig. 2. Daily density estimates with the corresponding bootstrap 95% confidence intervals.

In the end... just a truncated Poisson GLM with a single independent explanatory variable!


# Hands-on GLM example

A count regression – before you tried the Poisson  
(so... try the new tricks you just learned about!)

Environmental Modelling & Software 97 (2017) 112–129

---

Contents lists available at [ScienceDirect](#)

 Environmental Modelling & Software


journal homepage: [www.elsevier.com/locate/envsoft](http://www.elsevier.com/locate/envsoft)

---

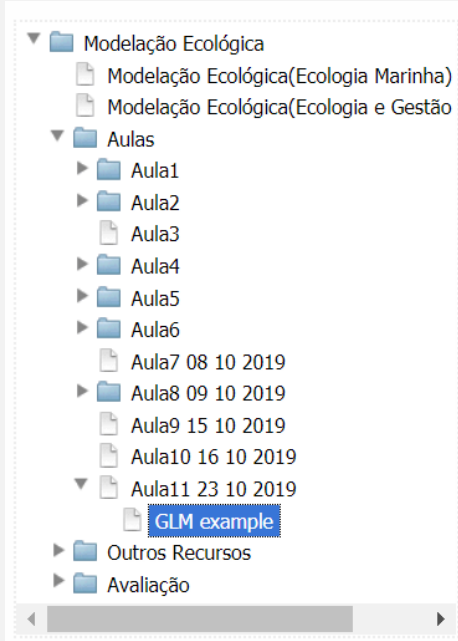
Application of random forest, generalised linear model and their hybrid methods with geostatistical techniques to count data: Predicting sponge species richness

Jin Li <sup>a,\*</sup>, Belinda Alvarez <sup>b,1</sup>, Justy Siwabessy <sup>a</sup>, Maggie Tran <sup>a</sup>, Zhi Huang <sup>a</sup>, Rachel Przeslawski <sup>a</sup>, Lynda Radke <sup>a</sup>, Floyd Howard <sup>a</sup>, Scott Nichol <sup>a</sup>

<sup>a</sup> Geoscience Australia, GPO Box 378, Canberra, ACT 2601, Australia  
<sup>b</sup> Museum and Art Gallery of the Northern Territory, PO Box 4646, Darwin, NT 0801, Australia

 CrossMark

Using the data in file “1-s2.0-S1364815217301615-mm2.csv” (FENIX folder “Count data GLM”) explain the variation in the response variable “sponge species richness” (species.richness) as a function of the other variables in said file – try everything but the Poisson!



### GLM example

Página **Ficheiros** 4 Permissões Link

Adicionar Ficheiro

#	Nome
1	1-s2.0-S1364815217301615-mm3.csv
2	1-s2.0-S1364815217301615-mm2.csv
3	1-s2.0-S1364815217301615-mm1.docx
4	1-s2.0-S1364815217301615-main.pdf

This data set is used in the paper below, feel free to explore the paper for details 😊, brief variable description in next slides

Environmental Modelling & Software 97 (2017) 112–129



Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Environmental Modelling & Software

journal homepage: [www.elsevier.com/locate/envsoft](http://www.elsevier.com/locate/envsoft)



Application of random forest, generalised linear model and their hybrid methods with geostatistical techniques to count data: Predicting sponge species richness

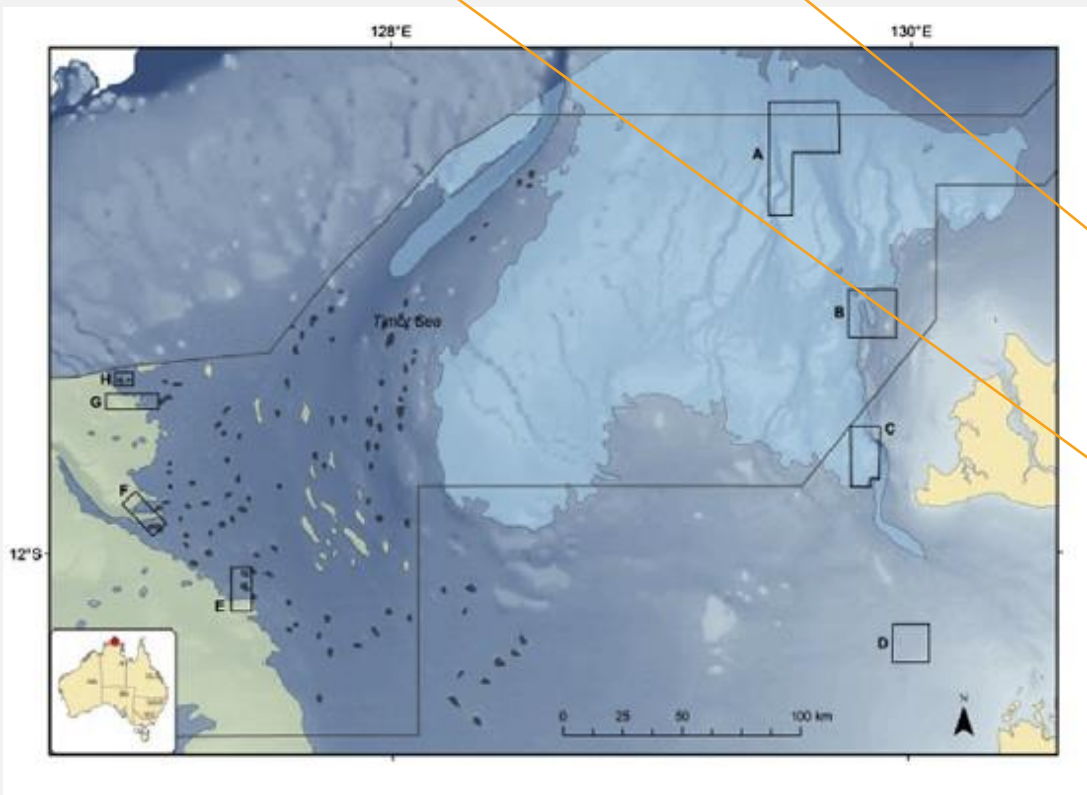
Jin Li <sup>a,\*</sup>, Belinda Alvarez <sup>b,1</sup>, Justy Siwabessy <sup>a</sup>, Maggie Tran <sup>a</sup>, Zhi Huang <sup>a</sup>, Rachel Przeslawski <sup>a</sup>, Lynda Radke <sup>a</sup>, Floyd Howard <sup>a</sup>, Scott Nichol <sup>a</sup>

<sup>a</sup> Geoscience Australia, GPO Box 378, Canberra, ACT 2601, Australia

<sup>b</sup> Museum and Art Gallery of the Northern Territory, PO Box 4646, Darwin, NT 0801, Australia



collection (Schlacher et al., 2007). There were 85 samples collected, and of which eight samples were excluded due to the uncertainty about transect length. In total, 77 samples were selected and used in this study. SSR is count data based on the presence/absence data, ranging from 1 to 39, with a mean of 10.48 and a standard deviation of 10.53. The point locations of samples are the mid-point of each transect.



```
> with(data,range(species.richness))  
[1] 1 39  
> #the range of the response variable  
> with(data,mean(species.richness))  
[1] 10.48052  
>  
> #the range of the response variable  
> with(data,sd(species.richness))  
[1] 10.52517
```

### 2.3. Predictive variables

Following a preliminary analysis based on data availability and the relationships with seabed hardness as discussed above and in previous studies, 80 predictive variables were available for this study. They are:

- 1) Two location variables: latitude (lat) and longitude (long),
- 2) Three sediment variables: mud, sand and gravel,
- 3) Bathymetry (bathy),
- 4) Twenty-seven backscatter (bs) variables (bs10 to bs36): a diffused reflection of acoustic energy due to scattering process back to the direction from which it's been generated, measured as the ratio of the acoustic energy sent to a seabed to that returned from the seabed, normalised to incidence angles between  $10^\circ$  and  $36^\circ$ ,
- 5) Seventeen derived variables from bs25 based on object and windows (30 m, 50 m and 70 m) approach:
  - a. bs\_o,
  - b. homogeneity (bs\_homo\_o, bs\_homo3, bs\_homo5, bs\_homo7),
  - c. entropy (bs\_entro\_o, bs\_entro3, bs\_entro5, bs\_entro7),
  - d. Local Moran I (bs\_lmi\_o, bs\_lmi3, bs\_lmi5, bs\_lmi7),
  - e. Variance (bs\_var\_o, bs\_var3, bs\_var5, bs\_var7).
- 6) Twenty-nine derived variables from bathy using object and windows (30 m, 50 m and 70 m) approach:
  - a. bathy\_o,
  - b. lmi\_o, lmi3, lmi5, lmi7,
  - c. Topographic position index (tpi\_o, tpi3, tpi5, tpi7),
  - d. Seabed slope (slope\_o, slope3, slope5, slope7),
  - e. Planar curvature (plan\_cur\_o, plan\_cur3, plan\_cur5, plan\_cur7),
  - f. Profile curvature (prof\_cur\_o, prof\_cur3, prof\_cur5, prof\_cur7),
  - g. Topographic relief (relief\_o, relief3, relief5, relief7),
  - h. Seabed rugosity (rugosity\_o, rugosity3, rugosity5, rugosity7).
- 7) Distance to coast (dist.coast)

